DOCUMENT RESUME

ED 482 927 TM 035 402

AUTHOR	Swygert, Kimberly A.
TITLE	The Relationship of Item-Level Response Times with Test-Taker and Item Variables in an Operational CAT Environment. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-CTR-98-10
PUB DATE	2003-10-00
NOTE	40p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Item Response Theory; *Participant Characteristics; *Responses; Test Format; *Test Items

ABSTRACT

In this study, data from an operational computerized adaptive test (CAT) were examined in order to gather information concerning item response times in a CAT environment. The CAT under study included multiple-choice items measuring verbal, quantitative, and analytical reasoning. The analyses included the fitting of regression models describing the relations of item parameters, ability measures, and item serial position with item-level response times. All of these analyses were performed for data collected under two different conditions: the first in which the test takers were only required to answer 80% of the items to receive a score, and the second in which the test-taker score was proportional to the number of items answered. The results for 2 datasets, 1 with 21,366 test takers and 1 with 11,301 test takers were compared. The results show that ability is predictive of item-level response times for items on the verbal section for both datasets, while item difficulty is predictive of item-level response times for certain sets of quantitative and analytical items. In each case, the regression equations explain more of the variability in the item-level response times when the data were administered under the proportional adjustment scoring rule. An appendix describes the method used and gives an example. (Contains 11 tables, 13 figures, and 29 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM TMTR 0252

ED 482 927

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. VASELECK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

LSAC RESEARCH REPORT SERIES

The Relationship of Item-level Response Times With Test-Taker and Item Variables in an Operational CAT Environment

Kimberly A. Swygert
Law School Admission Council

Law School Admission Council
Computerized Testing Report 98-10
October 2003

TM035402

A Publication of the Law School Admission Council

BEST COPY AVAILABLE

LSAC RESEARCH REPORT SERIES

The Relationship of Item-level Response Times With Test-Taker and Item Variables in an Operational CAT Environment

Kimberly A. Swygert
Law School Admission Council

Law School Admission Council
Computerized Testing Report 98-10
October 2003

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are 201 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

© 2003 by Law School Admission Council, Inc.

All rights reserved. No part of this report may be reproduced or transmitted in any part or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, Newtown, PA 18940-0040.

LSAT® and LSAC are registered marks of the Law School Admission Council, Inc.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary 1

Abstract 2

Introduction 2

Method 3
Data 3
Analysis of Item and Test-Taker Effects 4
Analysis of Rapid-Guessing Behavior 5

Results 6
Item Regressions 6
Test-Taker Regressions 8
Residual Regressions 10
Distributional Plots 16

Discussion 28
The Verbal Section 29
The Quantitative Section 29
The Analytical Reasoning Section 30
Additional Topics 30

References 31

Appendix 33

Executive Summary

The feasibility of implementing the Law School Admission Test (LSAT) as a computerized test (CT) or computerized-adaptive test (CAT) has been under investigation. One of the advantages of creating a computerized LSAT is that the item-level response times will become available for study; the paper-and-pencil (P&P) version does not allow for the practical collection of these response times. Issues such as speededness (when all test takers do not finish the test), differential speededness (when certain test-taker groups are more likely to be speeded than others, with groups differing in ways not related to the ability being measured), and the variability among test takers in response times are as important in a CAT format as in a P&P format; with item-level response times being available, these traditional issues may be studied in new ways.

It is now possible to look at test taker behavior by observing response times across the test; when test takers begin to run out of time, they often begin responding rapidly to the final items. One question that may be asked is whether the tendency to respond rapidly near the end of a test is independent of ability. Many existing studies based on operational and simulated computerized tests (CTs) assume that independence exists. Even if this assumption were true, it cannot be assumed that all variability in test-taker response times is independent of ability. It is possible that, for some items, speed is part of the underlying construct being measured (even if that was not the intention of the test developers). Studies that have examined the relationship between ability and response time suggest that, under unspeeded or untimed conditions, no relationship may be apparent, but when speededness is present, a relationship may materialize. Given that most computerized tests are administered under a time limit, and that a computerized LSAT may well be also, the relationship between ability and response time should be examined rather than assumed to be nonexistent.

Another important question involves the relationship of item characteristics to mean (average) item response time. Both item difficulty and item serial position within the test may be related to response time. If difficulty is predictive of response time, this may need to be taken into account when items are administered, so that test takers who are receiving difficult items are not handicapped by the time limit for the test. If item serial position is related to response time, that may be another measure of overall test speededness, particularly if the response times greatly decrease near the end of the test.

In this study, data from an operational CAT were examined in order to gather information concerning item response times in a CAT environment. The CAT included multiple-choice items measuring verbal, quantitative, and analytical reasoning; only discrete, stand-alone items were used, to avoid the confounding of item response time with reading time for passages. The analyses included the fitting of regression models describing how well the variability in the item-level response times is predicted by item-response-theory-based (IRT-based) item parameters and serial position. All of these analyses were performed for CAT data collected under two different conditions: the first, in which the test takers were only required to answer 80% of the items to receive a score, and the second, in which the test-taker score was proportional to the number of items answered. The availability of two versions of the same test, differing only by the number of items the test taker is required to answer, was a bonus for the study, as it was possible to compare the results for the two datasets to see if the change in scoring rule produced a change in the relationships among the variables.

One issue that arose with the use of response times as a variable in model fitting was the issue of estimating response time means. A CAT data array will have missing values in nonrandom ways, and this had to be taken into account when estimating the mean item response times. Godfrey's solution to mean estimation with missing data, the square combining table method, was employed to give better estimates of mean response times to be used in the fitting of the regression models. Also-to account for the relationship between ability and item difficulty-the effects due to items, the effects due to test takers, and the residual effects were all isolated, and the regressions were performed separately for each set of effects.

The results show that ability is predictive of item-level response time for items on the verbal section for both datasets, while item difficulty is predictive of item-level response times for certain sets of quantitative and analytical items. In each case, the regression equations explained more of the variability in the item-level response times when the data were administered under the proportional adjustment scoring rule, which produced a more speeded testing situation than did the 80% scoring rule. Thus, it appears that test takers with low verbal abilities may be more affected by the testing time limits, and this conclusion agrees with earlier research on differential speededness. Also consistent with earlier research are the conclusions that (1) more difficult quantitative items take longer to do, even if the test taker knows how to do them, (2) the analytical section is speeded for test takers of all abilities, and (3) despite the speededness of the analytical section, it appears that those items require more time as they become more difficult, and low-ability test takers may be working less quickly on this section than high-ability test takers. Due to using only discrete items in the analyses, the generalizability of these results to the LSAT is limited, as the LSAT contains no

quantitative section and contains only passage-based items on the reading comprehension and analytical reasoning sections. However, the results from the analytical reasoning section of this study do generalize to the logical reasoning section of the LSAT, because the item types are similar.

Abstract

In this study, data from an operational computerized-adaptive test (CAT) are examined in order to gather information concerning item response times in a CAT environment. The CAT under study includes multiple-choice items measuring verbal, quantitative, and analytical reasoning. The analyses include the fitting of regression models describing the relations of item parameters, ability measures, and item serial position with item-level response times. All of these analyses are performed for data collected under two different conditions: the first in which the test takers were only required to answer 80% of the items to receive a score, and the second in which the test-taker score was proportional to the number of items answered. The results for the two datasets are compared. The results show that ability is predictive of item-level response time for items on the verbal section for both datasets, while item difficulty is predictive of item-level response times for certain sets of quantitative and analytical items. In each case, the regression equations explain more of the variability in the item-level response times when the data were administered under the proportional adjustment scoring rule.

Introduction

The feasibility of a computerized test (CT) or computerized-adaptive test (CAT) version of the Law School Admission Test (LSAT) has been under investigation (Pashley, 1995). One of the advantages of creating a computerized LSAT is that the item-level response times will become available for study; the paper-and-pencil (P&P) version does not allow for the practical collection of these response times. Traditionally, response time studies have relied on overall test times, and these studies usually focus on the test-taker test-completion times, or the number of items a test taker completes in a given time period. Of the studies of test-level response times that exist, many involve the examination of test speededness, meaning the effect of time limits on test-taker test-completion behavior. Speeded test-taker behavior has been found in many large-scale standardized P&P tests, computerized tests (CTs) and CATs (Schaeffer, Reese, Steffen, McKinley, & Mills, 1993; Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995; Schnipke, 1995; Slater & Schaeffer, 1996), and has been found to occur differentially for different ethnic subgroups (Dorans, Schmitt, & Bleistein, 1988; Lawrence, 1993; Schmitt, Dorans, Crone, & Maneckshana, 1991). Eignor, Stocking, Way, and Steffen (1993) present the theory that, ideally, a CAT should be unspeeded, but while switching a test from P&P to CAT often reduces the number of items that must be administered to obtain a stable estimate of ability, the absence of speededness in the CAT format cannot be assumed.

While the existing body of literature on test-level response times provides valuable information about how the presence of an overall time limit affects test performance on the whole, these studies do not reveal how individual item and test-taker characteristics might affect test-taker response speed. The issue of the relationship of item-level response times with test-taker ability (also denoted as θ) and item characteristics (such as the item parameters of difficulty, discrimination, and guessing, among others) is an important one; in order to research this issue, studies that involve item-level response times are needed.

Many of the existing simulation and operational studies that examine the relationship between ability and item-level response times (Oshima, 1994; Schnipke, 1995; Schnipke & Scrams, 1996) have been concerned with *rapid-guessing behavior*, where the test taker begins to respond rapidly near the end of the test when time is running out (rapid-guessing behavior is in contrast to *solution behavior*, where the test taker takes sufficient time to solve the item). These studies have assumed that the tendency to run out of time, or to rapid-guess, is independent of ability. Even if this assumption were true, it cannot be assumed that all of the variability in test-taker response times is independent of ability. It is possible that, for some items, speed is part of the underlying construct being measured (even if that was not the intention of the test developers). Given that most computerized tests are administered under a time limit, and that a computerized LSAT would be as well, this relationship between ability and response time should be examined rather than assumed to be nonexistent.

Studies which have actually examined the relationship between test-taker ability and response times provide inconclusive results. One study, which used the computerized version of the National Board of Medical Examiners (NBME), found no correlation between test-taker ability and response times (Swanson, Featherman, Case, Leucht, & Nungester, 1997). However, when a replication was performed with the Step 1 Licensure Exam of the NBME, a positive correlation was discovered (van der Linden, Scrams, & Schnipke, 2002). The Step 1 exam is administered with a more stringent time limit than the general NBME, and so it is possible that there is no relationship between speed and accuracy on unspeeded tests, but that speed and accuracy are related on speeded tests.

A case in point is a study of the Graduate Record Examination computer-based test (GRE-CBT), which is a speeded test (Schaeffer et al., 1993). This study found support for a negative correlation between the ability being measured and response time. The researchers explored this issue by examining the average amount of time that high-, medium-, and low-ability test takers spent on the items. There appeared to be no difference among the ability groups when response times were averaged across all item difficulty levels, but differences did appear when the times for easy, medium, and difficult items were considered separately. The differences appeared at the extremes of ability and item difficulty: When the items were easy, high-ability test takers took less time than low-ability test takers, but when the items were difficult, low-ability test takers spent less time answering on average than the high-ability test takers. One explanation for the results is that once the items became too difficult for the low-ability test takers, those test takers may have started rapidly guessing. The results from this study suggest that not even the variability in response time that is due to rapid-guessing behavior is independent of ability.

Another study, this one using items from a CAT, found a relationship between item difficulty and average item response time. Van der Linden, et al. (2002) examined Arithmetic Reasoning items from the computerized-adaptive version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). The purpose of the paper was to examine the usefulness of a new item-selection algorithm in CAT to remove differential speededness. One side effect of this was the discovery that slow test takers of high ability might be differentially affected by time limits, as the more difficult arithmetic items required more time to solve. Given this, it is possible that a CAT arithmetic item pool might need to include several difficult items that have short response times, or, if arithmetic item difficulty is inherently linked to response time, it might be necessary to allow those of high ability to answer fewer items within the test's time limits.

In addition to item difficulty, Schnipke and Scrams (1996, 1997) suggest that an item's serial position in the test may be related to its average response time. They tested a two-state mixture model (proposed by Luce, 1986) that describes the response time distribution for an item using the proportion of rapid guessing on that item, the rapid guessing time distribution for that item, and the solution behavior response time distribution for that item. The two-state model was supported by the data, which came from an operational CT. The rapid guessing response time distribution appeared to be the same for all items; items on the last half of the test had many more rapid responses, making rapid-guessing behavior a function solely of item position.

In summary, the evidence suggests that certain item and test-taker variables may help explain some of the variability in item-level response times. Given that the current research plan for the LSAT is to develop a computerized and possibly adaptive form of the test, it is possible that the relationships of these variables with response times are strong enough to warrant taking response time into account when evaluating test-taker behavior, or when assembling items for a test (if the proposed assembly algorithm involves any measure of response time when selecting items).

The current study attempts to address these issues with the use of an operational CAT dataset. This CAT is composed of multiple-choice items measuring verbal, quantitative, and analytical reasoning. The analyses include the fitting of regression models describing the relations of item parameters, the test-taker ability measure, and item serial position with item-level response times. All of these analyses are performed for CAT data collected under two different conditions: the first, in which the test takers were only required to answer 80% of the items to receive a score, and the second, in which the test-taker score was proportional to the number of items answered. The availability of two versions of the same test, differing only by the number of items the test taker is required to answer, is a bonus for the study, as the results for the two datasets can be compared to see if the change in scoring rule produced a change in the relationships among the variables.

Method

Data

Two sets of operational CAT data are used in the current study. Each dataset contains three sections: A verbal section (consisting of 30 sentence completion, analogy, antonym and reading comprehension items), a quantitative section (containing 28 comparison, data interpretation, and discrete mathematical manipulation items), and an analytical reasoning section (which had 35 items that require logical as well as analytical reasoning). Although these are not LSAT items, the reading comprehension items are very similar to the LSAT Reading Comprehension (RC) items, and many of the analytical reasoning items are very similar to LSAT Analytical Reasoning (AR), and Logical Reasoning (LR) items. The items are multiple-choice, scored as correct or incorrect, and there is no penalty for guessing; these procedures are consistent with the LSAT item scoring procedures.

The first dataset (also referred to as Dataset 1) contains 21,366 test takers and was administered under an *80%-rule* scoring method. Once the time limit for each section had expired, a test taker would receive a scaled score (based on the θ-value that was estimated from their items response pattern) for that section only

if he or she had answered at least 80% of the items. However, the test takers could stop at 80% of the items and not answer any more, because they were aware that they would be graded as though they had answered the entire section. The second dataset (Dataset 2) is composed of 11,301 test takers and was administered under the same conditions as the first dataset, except that the 80% rule was replaced by the *proportional-adjustment* scoring method. This scoring rule gives all test takers a score based on their response patterns and on the number of items that they answered. The θ estimated from the item responses is multiplied by the proportion of items answered (out of the total number of items) before being converted to a scaled score; a test taker who does not answer every item cannot maximize his or her score. The datasets can be further divided into pools, four for Dataset 1 (pools 10, 11, 12, and 13) and three for Dataset 2 (pools 7, 8, and 9). Because the items and test takers are different for each of the seven pools, and because there were three sections to each test, each analysis that is subsequently mentioned was performed 21 times.

It should be noted that each item presented on this CAT is either a *discrete*, stand-alone item, or one item of a set. For discrete items, the response time recorded for each item began with the presentation of the item on the screen, and ended when the test taker clicked on a response. Set items are a related group of items all related to one passage. For set items, the passage associated with each item was displayed on the left-hand side of the screen, with the items appearing in turn on the right-hand side of the screen. The passage would remain on the left-hand side of the screen while items related to that passage continued to appear on the right-hand side. The result was that the response times for items that are related to a passage are confounded with the length of time the test taker spends reading or rereading the passage, and so all of the analyses that follow are done on discrete items only.

Analysis of Item and Test-taker Effects

The relationship of item-level response times with test-taker and item variables is examined with the use of a model that partitions the variability of response time[1] into several parts:

$$ln(t_{ij}) = m + v_i + w_j + r_{ij} + e_{ij}. \quad (1)$$

Here, $ln(t_{ij})$ is the natural log of response time of test taker j to item i, m is the grand mean of all the log response times, v is the effect of item i (i.e., change in response time due to the item), w is the effect of test taker j (change in response time due to the test taker), r_{ij} is the residual array of time after v and w have been removed, and e_{ij} is a normally distributed error component. Because of the number of variables in the data that are being regressed on response time, it is more tractable and informative to address the various relationships by dividing the response time data into three separate components (one item effect vector v, one test-taker effect vector w, and the residual array r), and regress the elements of these arrays on the pertinent variables, than it is to try to identify the effects of all the relevant variables with one overall regression. Separating the data into these components simplifies the problem, while also addressing the confounding of item and test-taker effects that occurs in a CAT administration.

The question is *how* to estimate these item and test-taker effects. If the test were linear, with all test takers responding to all items, the vector v of item effect estimates could be the column means across test takers of response times for each item, with the test-taker effect estimates w being the row means across items. However, a CAT test-taker-by-item data array is sparse, and more importantly, not randomly so, because every test taker does not see every item, and items are targeted as much as possible to each test taker's estimated θ level. The simple averages across items and test takers are not likely to be an adequate measure of those effects.

One possible solution is suggested by Godfrey (1985), who noted that information in a sparse two-way array could be arranged to form a less sparse representation with the use of a square combining table (SCT). The SCT method utilizes differences between pairs of values in the array to estimate row and column effects. For complete data, using the SCT method on a test-taker-by-item array of response times produces the same results as calculating the mean effects for items (v) with mean effects for test takers (w) already having been removed, and vice versa (an example of this on a small two-way array is provided in the Appendix).

---

[1]Most response time studies which involve any model-fitting tend to transform response times, as these are generally positively skewed. As the current analyses involved the fitting of least-squares regression models that assume normal error distributions, some transformation was required. Both logarithmic and square-root transformations were examined; the log transformation produced the most normal distribution of response times. These results are completely consistent with previous literature (Thissen, 1983; Schnipke & Scrams, 1997; Scrams & Schnipke, 1997) and so all subsequent references to "response time" in the current analyses mean log of response time (originally measured in seconds).

This method is likely to recover the mean effects when there are gaps in the data, but it is not completely resistant to missing data. It will fail to recover the effects if the square combining table itself contains gaps. In the context of the CAT data, this method is capable of producing adequate item and test-taker effect estimates only if every possible item pairing has been seen by at least one test taker, *and* every possible test-taker pair has at least one item in common. The first situation is plausible, because there are so many test takers; but because items on a CAT are assigned in part to be related to each test taker's ability, it is likely that there will be test-taker pairs, probably at opposite ends of the ability spectrum, who receive no items in common. Thus, for the items, the SCT process could be used to provide effect estimates that are an improvement over the simple means, but for the test-taker effects there is not necessarily any one method that is going to improve on using the mean as the estimate of the effect.

Therefore, the SCT method is used to produce the item effect estimates, these item effects are subtracted from the data, and test-taker means across the items are calculated to produce the test-taker effect estimates. The test-taker effects are then subtracted from the data, and what is left are the residual data points. Once the item, test-taker, and residual effects have been calculated, these effects are regressed on other variables. For the item effect regressions, the 3PL IRT-parameters a (discrimination), b (difficulty), and c (psuedo-guessing) are used as predictors. The test-taker effects are regressed on the ability estimate, θ. Finally, item serial position is used as a predictor for the residuals, in order to examine the pattern of response times across each section. The results of this third regression are used to examine whether Schnipke's (1995) claim that response times tend to decrease at the end of a linear timed test applies to an adaptive test as well.

Because item position includes information about the test taker (e.g., high-ability test takers are more likely to see difficult items at the end of the section than low-ability test takers) as well as about the item, it is possible that the coefficient for the third regression would interact with aspects of the test takers. The θ variable was categorized[2] and included into the residual regression equation, both as a main effect and as an interaction term with position, and the residual regressions are recalculated. Predicted values for item response times near the beginning and the end of the sections are calculated.

The fit of each model is assessed by the size and significance of the regression coefficients, and the percent of variability in the dependent measure explained by the independent measures. As mentioned above, each significance test is calculated separately for each pool, and so there are multiple tests of significance on each section within each dataset: For Dataset 1, there are 4 item regressions, 4 test-taker regressions, and 12 residual regressions; for Dataset 2, there are 3 item regressions, 3 test-taker regressions, and 9 residual regressions. The family-wise α-level is controlled using the Bonferroni method of dividing 0.05 by the number of tests: The Dataset 1 significance level is set at 0.0025, and the Dataset 2 significance level is set at 0.003. The sample sizes are so large for the test taker and the residual regressions that the F-tests are almost certain to be significant; thus, examination of the magnitude of the coefficients of determination (R2) is more useful.

Analysis of Rapid-Guessing Behavior

At this point, it is possible to examine the data further for rapid-guessing behavior, which is expected to be present. One method of measuring the amount of rapid-guessing behavior on an item is to examine the distributions of response times for that item separately for correct and incorrect responses. Rapidly guessed responses are likely to be incorrect, and so the distribution of response times for incorrect responses is likely to show a spike at the low response times, while the accompanying distribution of correct responses may not. Schnipke and Scrams (1997) examined this possibility by creating probability density function (PDF) plots of response time for correct vs. incorrect responses. They used CT items from a linear test, so they were able to make these plots for each item across all test takers. Some of their items appeared to elicit rapid-guessing behavior, and those items were more likely to occur at the end of a section.

While the PDF idea for visualizing rapid-guessing behavior is a good one, it needs to be modified for application to a CAT, where items occur in different positions for different test takers. Rapid-guessing behavior is more likely to appear for an item when it is administered in the last few serial positions than when it appears earlier in the section. Another way to say this is that near the end of the section, an item's position, rather than its content, is more likely to determine whether or not responses are speeded for a particular test taker, although other aspects of the item may have an effect as well.

There are two ways to deal with this situation. One way is to use response times to create PDFs for individual items. It is likely that a PDF graph for an item when it occurs in one of the last positions of the section would show speeded effects, but a PDF for the same item during the earlier positions would not. This

---

[2]Categorization involved assigning test takers to categories based on their estimated θ scores. The range of the θ variable was divided into six categories based on the areas under a normal distribution, so that the six categories all contained nearly the same sample size. This was done separately for each section of the test, because the distribution of θ is slightly different across the sections.

method is consistent with the use of the PDF plots for CT items, but it has two major drawbacks. One is that the CAT item selection algorithm uses constraining rules, so items are not administered to very many test takers. This means that the large sample sizes that Schnipke and Scrams obtained for each item would not be obtained with CAT data. The other drawback is that each CAT pool contains anywhere from 50 to 250 discrete items, so creating two PDFs for each item would produce more voluminous results than could easily be analyzed.

One alternative is to create PDFs not for individual items, but for serial positions. In this plan, one PDF is created for each of the verbal, quantitative, and analytical serial positions; both the sample size and pool problems would be solved. However, these graphs would be aggregating all the items that occurred in each position, and if there were aspects of the items that affected the likelihood of rapid-guessing behavior, those aspects would not be identifiable with this method. These effects would be visible with the first method (creating PDFs for each item), but, as only the very few items with large sample sizes can be graphed, the first method would not show if systematic differences in item content or characteristics affected rapid-guessing behavior.

For the current data, however, the use of the residual effects in place of response times, graphed for each position, would solve both problems. First, creating the PDF graphs for each position yields large sample sizes. Second, creating the PDF graphs using residual effect of time means that effects due to items and test takers are removed from the data. Therefore, if very short response times appear, they reflect rapid guessing, and the graphs reveal at what position these rapidly guessed responses begin. PDFs are graphed for residual effects for early, middle, and late positions for selected item pools from both of the datasets and for each of the three CAT sections.

Results

First, the data were examined for outliers, these defined as test takers who failed to answer more than five items in a section or who had average response times of less than 10 seconds per item on a section. Once outliers were removed, the final sample sizes were: Dataset 1—verbal section: 20,714; quantitative section: 20,657; analytical reasoning section: 20,512; Dataset 2—verbal section: 11,257; quantitative section: 11,252; analytical reasoning section: 11,221. In Dataset 1, 19,775 test takers answered at least 80% of the items on all three sections and can be said to have completed the entire test. The remaining test takers answered at least 80% of the items on either one or two sections.

Relatively commonly administered discrete items (those seen by at least 10 test takers) were selected for the fitting of the regression models. The resulting sets for the verbal section contained between 117 and 163 items per pool, those for the quantitative section contained between 181 and 223 items, and those for the analytical section—heavily dependent on set items—contained between 50 and 68 discrete items. The number of items per pool is denoted N_I, and the number of test takers exposed to each pool is denoted N_E.

The SCT method was used to calculate the item effects vector v, of dimension N_I. These effects were subtracted from the data array. From the resulting first residual array of the data, the test-taker mean vector w of dimension N_E was calculated. These test-taker effects were subtracted from the first residual array to produce the second residual array r, which was restructured to be a vector of size $N_I * N_E$. These three vectors v, w, and r were used as the dependent variables for regression analyses.

Item Regressions

The results for the item effect regressions are presented in Table 1 for Dataset 1 and Table 2 for Dataset 2. None of the item effects in Dataset 1 are significant at the Bonferroni-adjusted p-values. The results from analytical pools 12 and 13 have the largest coefficients of determination, which are 0.18 and 0.14.

TABLE 1

Item effect regression parameter estimates for Dataset 1: R^2, F, p, and β coefficients and standard errors

	Pool 10		Pool 11		Pool 12		Pool 13	
	Value	SE	Value	SE	Value	SE	Value	SE
N_i	123		131		129		117	
Verbal								
R^2	0.02		0.01		0.05		0.02	
F	0.79		0.64		2.30		0.83	
p	0.50		0.59		0.08		0.48	
a	-0.20	0.14	-0.17	0.13	0.09	0.16	-0.07	0.11
b	0.02	0.03	0.01	0.03	-0.03	0.03	0.05	0.03
c	0.03	0.33	0.11	0.34	-0.82	0.38	-0.21	0.35
N_i	187		187		194		181	
Quantitative								
R^2	0.01		0.01		0.002		0.03	
F	0.47		0.92		0.15		1.51	
p	0.71		0.43		0.93		0.21	
a	-0.03	0.13	0.14	0.13	-0.01	0.14	-0.07	0.11
b	0.05	0.04	-0.04	0.04	0.01	0.04	-0.05	0.04
c	-0.02	0.34	-0.45	0.34	0.14	0.27	-0.20	0.36
N_i	54		55		50		56	
Analytical								
R^2	0.01		0.05		**0.18**		**0.14**	
F	0.19		0.89		**3.45**		**2.89**	
p	0.90		0.45		**0.02**		**0.04**	
a	-0.86	0.16	-0.20	0.23	-0.32	0.22	-0.25	0.14
b	0.01	0.03	0.01	0.03	-0.05	0.03	0.06	0.06
c	0.23	0.40	0.69	0.44	**1.34**	0.43	0.45	0.21

Note. N_i = Number of discrete items. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted *p*-value of .004.

TABLE 2

Item effect regression parameter estimates for Dataset 2: R^2, F, p, and β coefficients and standard error

	Pool 7		Pool 8		Pool 9	
	Value	SE	Value	SE	Value	SE
N_i	163		161		163	
Verbal						
R^2	**0.08**		**0.08**		**0.06**	
F	**4.90**		**4.36**		**3.30**	
p	**0.003**		0.006		**0.02**	
a	0.02	0.11	0.11	0.14	-0.02	0.13
b	**0.09**	0.03	**0.09**	0.03	**0.08**	0.03
c	**-0.80**	0.36	-0.55	0.31	-0.30	0.33
N_i	216		237		223	
Quantitative						
R^2	**0.10***		**0.16***		**0.21***	
F	**7.51**		**15.25**		**19.67**	
p	**0.0001**		**0.0001**		**0.0001**	
a	-0.02	0.10	-0.05	0.12	-0.18	0.11
b	**0.14***	0.04	**0.19***	0.03	**0.23***	0.03
c	-0.04	0.32	-0.33	0.30	-0.16	0.27
N_i	67		68		65	
Analytical						
R^2	**0.26***		**0.21***		**0.13**	
F	**7.51**		**5.83**		**3.14**	
p	**0.0002**		**0.001**		**0.032**	
a	**1.03**	0.93	**1.25**	0.55	1.08	0.67
b	**0.13***	0.04	**0.12***	0.03	0.06	0.04
c	-0.52	0.42	**-1.08**	0.52	-0.78	0.65

Note. N_i = Number of discrete items. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted *p*-value of .006.

The item regression results for Dataset 2 are very dissimilar from those for Dataset 1. The coefficient of determination R_2 for the overall item parameter effect is significantly different from zero for every quantitative pool and two of the three analytical pools. While the verbal coefficients of determination are not significant at the Bonferroni-adjusted α-level, they are larger than the values in Dataset 1. Out of the five significant regressions for Dataset 2, the coefficient for the *b* parameter is significant in all of them, that for *a* is significant twice (analytical pool 7 and Pool 8), and the coefficient for *c* is significant once (analytical pool 8). The positive coefficients for *b* for the quantitative and analytical sections indicate that for Dataset 2, but not Dataset 1, difficult items tend to take longer.

It is not immediately clear why these results would appear for one dataset but not the other. The sample size N_I is larger in Dataset 2, but it does not appear to be so much larger as to explain these effects. One possible explanation is that the distribution of the item parameters in these pools could be very different across pools (and across the two datasets). It is known that there was no attempt to equate this distribution across pools, and an artificially restricted range of the item parameters in one or more pools could have occurred, which would reduce the predictive power of these parameters.

A MANOVA with the three item parameters as dependent variables was performed for each section to see if the centroid of the trivariate item parameter distribution differed across pools. The results of the three MANOVAs are printed in Table 3. The α-level was set at 0.05 and then divided by 3 to produce an adjusted α-level of 0.017. None of the effects were significant across the seven pools (verbal $p < 0.18$; quantitative $p < 0.08$, and analytical $p < 0.04$). It does not appear that the seven pools in each section have different distributions of item parameters. Given the results of the item effect regression and MANOVA analyses, it appears that something is different between Dataset 1 and Dataset 2 that cannot be explained by the distributions of the item parameters. The only procedural difference acknowledged by the testing agency is the 80% rule used for scoring in Dataset 1, but not in Dataset 2.

TABLE 3
MANOVA for the test of the equality of the trivariate a, b, and c distributions across pools

	Pool 7 Means	Pool 8 Means	Pool 9 Means	Pool 10 Means	Pool 11 Means	Pool 12 Means	Pool 13 Means
Verbal							
a	0.63	0.59	0.60	0.65	0.58	0.64	0.65
b	-0.37	-0.41	-0.37	-0.49	-0.51	-0.54	-0.51
c	0.15	0.17	0.17	0.16	0.18	0.16	0.14
	Equality of centroids: $F_{(18,1149)} = 1.30$ $p < 0.18$; $\eta = .23$						
Quantitative							
a	0.63	0.59	0.60	0.65	0.58	0.64	0.65
b	0.07	0.03	0.12	0.13	0.19	0.18	0.03
c	0.14	0.15	0.15	0.15	0.17	0.17	0.13
	Equality of centroids: $F_{(18,1149)} = 1.52$ $p < 0.08$; $\eta = .25$						
Analytical							
a	0.63	0.59	0.60	0.65	0.58	0.64	0.65
b	0.06	0.25	-0.03	0.44	-0.06	0.10	0.06
c	0.16	0.13	0.17	0.15	0.13	0.14	0.16
	Equality of centroids: $F_{(18,1149)} = 1.67$ $p < 0.04$; $\eta = .27$						

Note. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted *p*-value of .017.

Test-Taker Regressions

The test-taker regressions are presented in Tables 4 and 5 for the two datasets. Here, more consistency is seen for the verbal section across the datasets than was seen for the item effect regressions; prediction of w by θ is significant for each pool. The coefficients of determination, however, differ more between the two datasets than within. For Dataset 1 the four verbal coefficients of determination are 0.16, 0.17, 0.22, and 0.14. In contrast, the Dataset 2 coefficients of determination are all larger than the ones in Dataset 1; the values are 0.28, 0.33, and 0.23 for the three pools. For the verbal pools, every coefficient for θ is negative for all datasets and pools, indicating that the more proficient test takers tend to take less time on the items.

TABLE 4
Test-taker effect regression parameter estimates for Dataset 1: R^2, F, p, and β coefficients and standard errors

	Pool 10		Pool 11		Pool 12		Pool 13	
	Value	SE	Value	SE	Value	SE	Value	SE
N_E	7,037		4,959		4,152		4,546	
Verbal								
R^2	**0.16***		**0.17***		**0.22***		**0.14***	
F	**1,399.40**		**1,006.60**		**1,147.50**		**757.88**	
p	**0.0001**		**0.0001**		**0.0001**		**0.0001**	
θ	**-0.10**	0.003	**-0.11**	0.003	**-0.11**	0.003	**-0.10**	0.004
N_E	6,973		4,951		4,148		4,546	
Quantitative								
R^2	**0.05***		**0.01***		**0.06***		**0.01***	
F	**336.30**		**72.70**		**242.46**		**43.58**	
p	**0.0001**		**0.0001**		**0.0001**		**0.0001**	
θ	**0.05**	0.002	**0.03**	0.003	**0.06**	0.004	**0.02**	0.003
N_E	6,796		4,951		4,141		4,528	
Analytical								
R^2	**0.03***		**0.03***		0.00		**0.05***	
F	**204.33**		**133.20**		5.41		**231.56**	
p	**0.0001**		**0.0001**		0.02		**0.0001**	
θ	**-0.05**	0.003	**-0.05**	0.004	-0.01	0.004	**-0.06**	0.004

Note. N_E = Number of test takers. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted p-value of .0025

TABLE 5
Test-taker effect regression parameter estimates for Dataset 2: R^2, F, p, and β and standard errors

	Pool 7		Pool 8		Pool 9	
	Value	SE	Value	SE	Value	SE
N_E	4,123		3,910		3,224	
Verbal						
R^2	**0.28***		**0.33***		**0.23***	
F	**1,542**		**1,962**		**954.9**	
p	**0.0001**		**0.0001**		**0.0001**	
θ	**-0.13**	0.003	**-0.15**	0.003	**-0.11**	0.004
N_E	4,118		3,910		3,224	
Quantitative						
R^2	**0.005***		0.00		0.00	
F	**18.83**		0.76		0.68	
p	**0.0001**		0.39		0.41	
θ	**-0.015**	0.004	-0.003	0.003	-0.003	0.004
N_E	4,087		3,903		3,224	
Analytical						
R^2	**0.13***		**0.11***		**0.08***	
F	**584.6**		**498.66**		**260.8**	
p	**0.0001**		**0.0001**		**0.001**	
θ	**-0.11**	0.004	**-0.1**	0.004	**-0.07**	0.004

Note. N_E = Number of test takers. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted p-value of .003.

The quantitative coefficients of determination are significant in all pools for Dataset 1, but are significant in only one pool of Dataset 2. The significant coefficients of determination are much smaller than the verbal values; R^2 = 0.05, 0.01, 0.06, and 0.01 in Dataset 1, and R^2 = 0.005 in Dataset 2. The coefficient for θ is positive for Dataset 1, indicating that the more proficient test takers are taking more time on the items. In Dataset 2, this coefficient becomes negative, indicating that the higher proficient test takers now take less time on the items. However, the absolute values of the R^2's are so small as to be practically meaningless.

The analytical reasoning results are more similar to the verbal results than to the quantitative results. The analytical coefficients of determination range from 0.00 to 0.05 for Dataset 1, and from 0.08 to 0.13 in Dataset 2. For the analytical sections, the coefficients for θ are always negative. There may be an effect on the analytical section, similar to the one on the verbal section, in which test takers who score higher work more quickly on the items, and this effect increases in Dataset 2. All of the test takers in Dataset 2 would have felt pressure to work quickly because they would be penalized for all items left unanswered, and so the presence

of the proportional-adjustment rule does not explain why the verbal and analytical high scorers would be working more quickly than the low scorers. One explanation is that test takers who are proficient in these sections can answer items at their level of difficulty more quickly—speed may be part of the construct. If this is the case, it would make sense for the effect to appear more pronounced when the scoring rule is switched from one with a less extreme time constraint (one only has to answer 80% of the items) to one with a more extreme time constraint (one should try to answer all the items).

Residual Regressions

The residual regression results are presented in several tables. First, Tables 6 and Table 7 present the results of regressing the residuals on item serial position only for Datasets 1 and 2 respectively. Tables 8 and 9 present the R^2 values and the F-values for the change in R^2 for the additional regressions of (*a*) residual effects on position and θ category, and (*b*) residual effects on position, θ category, and θ-by-position interactions. Finally, Tables 10–15 present the regression coefficients along with selected predicted residual values for a model with position, θ category, and θ-by-position interaction terms (Tables 10–11 represent the verbal section, Tables 12–13 represent the quantitative section, and Tables 14–15 represent the analytical reasoning section).

Tables 6 and 7 show that all of the coefficients of determination for position are significant for both datasets and all sections. This is due in part to the sample size ($N_I{}^*{}_E$); almost any F-value calculated with a six-digit sample size is significant. None of the coefficients of determination are greater than 0.07; an item's position alone does not predict a great deal of the variability in the residual effect of time. The standard errors are extremely small, however, so the descriptions of the relation between item position and residual log response time, eliminating test-taker and item effects, are very accurate. The coefficients of determination are also consistent across datasets; the verbal and quantitative values range between 0.001 and 0.009, and the analytical values range between 0.02 and 0.09.

TABLE 6

Residual effect regression parameter estimates for Dataset 1: R^2, F, p, and β coefficients and standard errors

	Pool 10		Pool 11		Pool 12		Pool 13	
	Value	SE	Value	SE	Value	SE	Value	SE
N_{PE}	124,083		88,731		80,319		78,770	
Verbal								
R^2	**0.009***		**0.005***		**0.006***		**0.004***	
F	**1,138.65**		**452.41**		**519.01**		**285.15**	
p	**0.0001**		**0.0001**		**0.0001**		**0.0001**	
position	**-0.005**	0.0001	**-0.004**	0.0002	**-0.004**	0.0002	**-0.003**	0.0002
N_{PE}	89,803		107,913		151,733		98,913	
Quantitative								
R^2	**0.003***		**0.003***		**0.001***		**0.003***	
F	**422.54**		**311.56**		**134.15**		**268.36**	
p	**0.0001**		**0.0001**		**0.0001**		**0.0001**	
position	**-0.004**	0.0002	**-0.004**	0.0002	**-0.003**	0.0002	**-0.003**	0.0002
N_{PE}	48,425		35,101		28,940		34,568	
Analytical								
R^2	**0.03***		**0.04***		**0.02***		**0.04***	
F	**1,537.99**		**1,431.87**		**581.93**		**1,281.35**	
p	**0.0001**		**0.0001**		**0.0001**		**0.0001**	
position	**-0.001**	0.0002	**-0.01**	0.0002	**-0.001**	0.0002	**-0.01**	0.0002

Note. N_{PE} = The product of the number of test takers and the number of discrete items that were administered. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted p-value of .0025.

TABLE 7
Residual effect regression parameter estimates for Dataset 2: R^2, F, p, and β coefficients and standard errors

	Pool 7		Pool 8		Pool 9	
	Value	SE	Value	SE	Value	SE
N_{PE}	89,647		84,998		70,031	
Verbal						
R^2	**0.007***		**0.006***		**0.006***	
F	**596.65**		**474.35**		**439.79**	
p	**0.0001**		**0.0001**		**0.0001**	
position	**-0.004**	0.0002	**-0.004**	0.0001	**-0.004**	0.0002
N_{PE}	104,608		82,670		76,448	
Quantitative						
R^2	**0.006***		**0.006***		**0.003***	
F	**639.42**		**561.33**		**221.07**	
p	**0.0001**		**0.0001**		**0.0001**	
position	**-0.005**	0.0002	**-0.005**	0.0002	**-0.004**	0.0002
N_{PE}	34,822		33,286		27,600	
Analytical						
R^2	**0.07***		**0.09***		**0.04***	
F	**2,782.99**		**3,418.87**		**1,260.02**	
p	**0.0001**		**0.0001**		**0.0001**	
position	**-0.02**	0.0002	**-0.02**	0.0003	**-0.01**	0.0002

Note. N_{PE} = The product of the number of test takers and the number of discrete items that were administered. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted p-value of .003.

The addition of the θ-category variable (shown in the "position + θ" columns of Tables 8 and 9) does not improve prediction of the residuals very much. On the quantitative section of Dataset 1, the addition of this variable increases the coefficient of determination by about 0.008 on each pool. The coefficients do not change at all for the verbal and analytical sections. In Dataset 2, even less improvement is seen by the addition of the θ category; there is only one non-zero F-value. All of the non-zero F-values are significant at $p < 0.0001$, due mainly to the large sample size.

TABLE 8
The R^2 and F-values for change in R^2 for the residual effect regressions when θ category and interaction terms are added to the model (Dataset 1)

	Position	Position + θ		Position + θ + Interaction	
	R^2	R^2	F	R^2	F
Verbal					
Pool 10 N_{PE} = 124,082	0.009	0.009	0.00	**0.012***	**376.73**
Pool 11 N_{PE} = 89,108	0.005	0.005	0.00	**0.008***	**269.44**
Pool 12 N_{PE} = 88,422	0.006	**0.008***	**178.26**	**0.009***	**89.21**
Pool 13 N_{PE} = 78,768	0.004	0.004	0.00	**0.005***	**79.15**
Quantitative					
Pool 10 N_{PE} = 151,732	0.003	**0.01***	**1,517.25**	0.01	0.00
Pool 11 N_{PE} = 107,912	0.003	**0.009***	**217.77**	**0.01***	**108.99**
Pool 12 N_{PE} = 89,802	0.001	**0.011***	**453.97**	**0.014***	**273.20**
Pool 13 N_{PE} = 98,912	0.003	**0.001***	**79.203**	**0.003***	**198.40**
Analytical					
Pool 10 N_{PE} =48,325	0.031	0.031	0.00	**0.034***	**150.04**
Pool 11 N_{PE} = 35,101	0.039	0.039	0.00	**0.040***	**36.55**
Pool 12 N_{PE} = 28,940	0.020	0.020	0.00	**0.025***	**148.35**
Pool 13 N_{PE} = 34,568	0.036	0.036	0.00	**0.038***	**71.84**

Note. N_{PE} = The product of the number of test takers and the number of discrete items that were administered. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted p-value of .0025.

TABLE 9
The R^2 and F-values for change in R^2 for the residual effect regressions when θ category and interaction terms are added to the model (Dataset 2)

	Position	Position + θ		Position + θ + Interaction	
	R^2	R^2	F	R^2	F
Verbal					
Pool 07 N_{PT} = 89,647	0.007	0.007	0.00	**0.013***	**544.89**
Pool 08 N_{PT} = 84,998	0.006	0.006	0.00	**0.009***	**257.27**
Pool 09 N_{PT} = 70,031	0.006	0.006	0.00	**0.011***	**353.99**
Quantitative					
Pool 07 N_{PT} = 96,514	0.006	0.006	0.00	**0.008***	**194.56**
Pool 08 N_{PT} = 92,669	0.006	0.006	0.00	**0.008***	**186.81**
Pool 09 N_{PT} = 76,448	0.003	0.003	0.00	**0.005***	**153.64**
Analytical					
Pool 07 N_{PT} = 34,822	0.074	0.074	0.00	**0.077***	**113.14**
Pool 08 N_{PT} = 33,286	0.093	0.093	0.00	**0.096***	**110.42**
Pool 09 N_{PT} = 27,601	0.043	**0.044***	**28.86**	**0.045***	**28.89**

Note. N_{PT} = The product of the number of test takers and the number of discrete items that they were administered. Bold estimates are significant at $p < .05$ or smaller. Starred (*) estimates are significant at the Bonferroni-adjusted p-value of .003.

When the interaction terms are added to the prediction equations (shown in the "position + θ + interaction" columns), the coefficients increase by 0.001 to 0.005. The increases in R^2 are larger for the verbal section and smaller for the analytical section. These results can be interpreted as follows: Although the small size of the coefficients of determination means that there remains a great deal of variability in the residual effect of time to be explained, position, θ and their interaction explain a small portion of the variability consistently. The verbal section shows the highest increase in R^2 when the interaction terms are added, and so one would expect that the residual values for serial position would vary more across test takers of different θ levels than they would on the quantitative and analytical sections.

This explanation is illustrated by computing predicted residual values for all three sections of the test. The intercept values and coefficients for the two main effects and the interaction effect are listed in Tables 10–15. The first column in each table contains the θ levels, lowest to highest. The second column contains the intercept value for the regression equation when all the terms are included in the model. The position effect is constant across rows, and the θ effect and position-by-θ interactions each contribute five non-zero terms, due to the dummy coding of the six-level θ-category variable; there are 11 terms in the full model. The third, fourth, and fifth columns contain the coefficients for the position, θ, and position-by-θ effects.

TABLE 10
The residual regression coefficients for the verbal section of Dataset 1, with predicted values for items at positions 4 and 30

θ Level	Intercept	Position Coefficient	θ Coefficient	Position* θ Coefficient	Position 4 Value	Position 30 Value
Pool 10						
1	-0.00002	0.0002	0.0025	-0.0107	-0.039	-0.312
2	-0.00002	0.0002	0.0023	-0.0065	-0.022	-0.186
3	-0.00002	0.0002	-0.0002	-0.0051	-0.019	-0.147
4	-0.00002	0.0002	-0.0012	-0.0036	-0.014	-0.103
5	-0.00002	0.0002	-0.0010	-0.0036	-0.001	-0.004
6	-0.00002	0.0002	0.0000	0.0000	0.001	0.005
Pool 11						
1	-0.00007	0.0003	0.0040	-0.0098	-0.034	-0.281
2	-0.00007	0.0003	-0.0009	-0.0074	-0.029	-0.214
3	-0.00007	0.0003	-0.0014	-0.0050	-0.020	-0.142
4	-0.00007	0.0003	-0.0004	-0.0030	-0.011	-0.081
5	-0.00007	0.0003	0.0002	-0.0024	-0.008	-0.062
6	-0.00007	0.0003	0.0000	0.0000	0.001	0.008
Pool 12						
1	-0.0173	-0.0012	0.04456	-0.0064	-0.003	-0.200
2	-0.0173	-0.0012	0.03558	-0.0060	-0.011	-0.197
3	-0.0173	-0.0012	0.02832	-0.0034	-0.007	-0.126
4	-0.0173	-0.0012	0.02068	-0.0023	-0.011	-0.102
5	-0.0173	-0.0012	0.01717	-0.0038	-0.020	-0.150
6	-0.0173	-0.0012	0.00000	0.0000	-0.022	-0.053
Pool 13						
1	0.0003	0.0005	0.0013	-0.0082	-0.029	-0.229
2	0.0003	0.0005	0.0009	-0.0046	-0.015	-0.122
3	0.0003	0.0005	0.0008	-0.0046	-0.015	-0.122
4	0.0003	0.0005	-0.0003	-0.0047	-0.017	-0.126
5	0.0003	0.0005	-0.0005	-0.0021	-0.007	-0.048
6	0.0003	0.0005	0.0000	0.0000	0.002	0.015

TABLE 11
The residual regression coefficients for the verbal section of Dataset 2, with predicted values for items at positions 4 and 30

θ Level	Intercept	Position Coefficient	θ Coefficient	Position* θ Coefficient	Position 4 Value	Position 30 Value
Pool 7						
1	-0.00001	0.0013	-0.00013	-0.0121	-0.043	-0.324
2	-0.00001	0.0013	-0.00006	-0.0075	-0.025	-0.186
3	-0.00001	0.0013	-0.00009	-0.0060	-0.019	-0.141
4	-0.00001	0.0013	0.00006	-0.0035	-0.009	-0.066
5	-0.00001	0.0013	0.00005	-0.0017	-0.002	-0.012
6	-0.00001	0.0013	0.00000	0.0000	0.005	0.039
Pool 8						
1	0.00005	0.00017	0.0010	-0.0098	-0.037	-0.288
2	0.00005	0.00017	0.0015	-0.0054	-0.019	-0.155
3	0.00005	0.00017	0.0005	-0.0024	-0.008	-0.066
4	0.00005	0.00017	-0.0005	-0.0038	-0.015	-0.109
5	0.00005	0.00017	-0.0005	-0.0020	-0.008	-0.055
6	0.00005	0.00017	0.0000	0.0000	0.0001	0.005
Pool 9						
1	-0.0001	-0.0003	0.00048	-0.0100	-0.041	-0.309
2	-0.0001	-0.0003	0.00049	-0.0052	-0.022	-0.165
3	-0.0001	-0.0003	0.00044	-0.0044	-0.019	-0.141
4	-0.0001	-0.0003	0.00017	-0.0023	-0.010	-0.078
5	-0.0001	-0.0003	0.00004	-0.0005	-0.003	-0.024
6	-0.0001	-0.0003	0.00000	0.0000	-0.001	-0.009

TABLE 12
The residual regression coefficients for the quantitative section of Dataset 1, with predicted values for items at positions 4 and 28

θ Level	Intercept	Position Coefficient	θ Coefficient	Position* θ Coefficient	Position 4 Value	Position 28 Value
Pool 10						
1	-0.0011	-0.0023	0.0030	-0.0059	-0.031	-0.228
2	-0.0011	-0.0023	0.0015	-0.0007	-0.015	-0.084
3	-0.0011	-0.0023	0.0015	-0.0005	-0.011	-0.078
4	-0.0011	-0.0023	0.0012	-0.0017	-0.016	-0.112
5	-0.0011	-0.0023	0.0008	0.0001	-0.009	-0.062
6	-0.0011	-0.0023	0.0000	0.0000	-0.010	-0.066
Pool 11						
1	0.0003	-0.0036	-0.0017	-0.0031	-0.028	0.000
2	0.0003	-0.0036	-0.0009	0.00005	-0.015	-0.100
3	0.0003	-0.0036	-0.0004	0.00099	-0.011	-0.073
4	0.0003	-0.0036	0.0004	0.0013	-0.009	-0.064
5	0.0003	-0.0036	0.0001	-0.0002	-0.015	-0.106
6	0.0003	-0.0036	0.0000	0.0000	-0.014	-0.101
Pool 12						
1	-0.0013	-0.0048	0.0022	-0.0007	-0.021	-0.153
2	-0.0013	-0.0048	0.0015	0.0019	-0.011	-0.081
3	-0.0013	-0.0048	0.0013	0.0047	0.000	-0.003
4	-0.0013	-0.0048	0.0012	0.0040	-0.003	-0.023
5	-0.0013	-0.0048	0.0012	0.0015	-0.013	-0.093
6	-0.0013	-0.0048	0.0000	0.0000	-0.021	-0.136
Pool 13						
1	-0.0003	-0.0027	0.0009	-0.0039	-0.026	-0.184
2	-0.0003	-0.0027	0.0004	0.0005	-0.009	-0.062
3	-0.0003	-0.0027	0.0005	0.0010	-0.007	-0.047
4	-0.0003	-0.0027	0.0003	-0.00089	-0.014	-0.101
5	-0.0003	-0.0027	-0.0004	-0.00222	-0.020	-0.139
6	-0.0003	-0.0027	0.0000	0.0000	-0.011	-0.076

TABLE 13
The residual regression coefficients for the quantitative section of Dataset 2, with predicted values for items at positions 4 and 28

θ Level	Intercept	Position Coefficient	θ Coefficient	Position* θ Coefficient	Position 4 Value	Position 28 Value
Pool 7						
1	-0.0008	-0.0083	-0.0019	0.0025	-0.026	-0.165
2	-0.0008	-0.0083	0.0006	0.0043	-0.016	-0.112
3	-0.0008	-0.0083	0.0015	0.0061	-0.008	-0.061
4	-0.0008	-0.0083	0.0009	0.0048	-0.014	-0.098
5	-0.0008	-0.0083	0.0006	-0.0014	-0.039	-0.272
6	-0.0008	-0.0083	0.0000	0.0000	-0.034	-0.233
Pool 8						
1	-0.0005	-0.0081	0.0009	0.002	-0.024	0.000
2	-0.0005	-0.0081	0.0008	0.004	-0.016	-0.115
3	-0.0005	-0.0081	0.0005	0.007	-0.004	-0.031
4	-0.0005	-0.0081	0.0004	0.005	-0.013	-0.087
5	-0.0005	-0.0081	-0.0007	-0.002	-0.042	-0.284
6	-0.0005	-0.0081	0.0000	0.000	-0.033	-0.227
Pool 9						
1	-0.0015	-0.0100	0.0018	0.0042	-0.023	-0.162
2	-0.0015	-0.0100	0.0016	0.0084	-0.006	-0.045
3	-0.0015	-0.0100	0.0016	0.0090	-0.004	-0.028
4	-0.0015	-0.0100	0.0014	0.0071	-0.012	-0.081
5	-0.0015	-0.0100	0.0015	0.0089	-0.004	-0.031
6	-0.0015	-0.0100	0.0000	0.0000	-0.042	-0.282

TABLE 14
The residual regression coefficients for the analytical section of Dataset 1, with predicted values for items at positions 4 and 35

θ Level	Intercept	Position Coefficient	θ Coefficient	Position* θ Coefficient	Position 4 Value	Position 35 Value
Pool 10						
1	-0.0019	-0.0016	0.0109	-0.0090	-0.033	-0.362
2	-0.0019	-0.0016	0.0084	-0.0070	-0.027	-0.294
3	-0.0019	-0.0016	0.0059	-0.0090	-0.038	-0.367
4	-0.0019	-0.0016	-0.0013	-0.0100	-0.049	-0.409
5	-0.0019	-0.0016	-0.0073	-0.0077	-0.046	-0.334
6	-0.0019	-0.0016	0.0000	0.0000	-0.008	-0.057
Pool 11						
1	-0.0007	-0.0067	0.0051	-0.0042	-0.039	-0.235
2	-0.0007	-0.0067	0.0070	-0.0057	-0.043	-0.428
3	-0.0007	-0.0067	-0.00004	-0.0038	-0.043	-0.368
4	-0.0007	-0.0067	-0.00214	-0.0028	-0.041	-0.335
5	-0.0007	-0.0067	-0.00302	-0.0043	-0.048	-0.389
6	-0.0007	-0.0067	0.00000	0.0000	-0.028	-0.235
Pool 12						
1	-0.00002	-0.00001	0.0069	-0.0106	-0.036	-0.654
2	-0.00002	-0.00001	0.0022	-0.0064	-0.023	-0.222
3	-0.00002	-0.00001	0.0046	-0.0069	-0.023	-0.237
4	-0.00002	-0.00001	-0.0026	-0.0063	-0.028	-0.223
5	-0.00002	-0.00001	-0.0065	-0.0083	-0.039	-0.297
6	-0.00002	-0.00001	0.0000	0.0000	0.000	0.000
Pool 13						
1	0.0012	-0.0067	0.0058	-0.0039	-0.035	-0.364
2	0.0012	-0.0067	-0.0029	-0.0007	-0.031	-0.261
3	0.0012	-0.0067	-0.0053	-0.0004	-0.033	-0.253
4	0.0012	-0.0067	-0.0031	-0.0043	-0.046	-0.253
5	0.0012	-0.0067	0.0003	-0.0048	-0.045	-0.401
6	0.0012	-0.0067	0.0000	0.0000	-0.025	-0.233

TABLE 15
The residual regression coefficients for the analytical section of Dataset 2, with predicted values for items at positions 4 and 35

θ Level	Intercept	Position Coefficient	θ Coefficient	Position* θ Coefficient	Position 4 Value	Position 35 Value
Pool 7						
1	-0.0044	-0.0158	0.0087	0.0033	-0.046	-0.433
2	-0.0044	-0.0158	0.0068	0.0054	-0.039	-0.362
3	-0.0044	-0.0158	0.0048	0.0018	-0.056	-0.489
4	-0.0044	-0.0158	0.0039	-0.0029	-0.075	-0.655
5	-0.0044	-0.0158	0.0002	-0.0017	-0.074	-0.617
6	-0.0044	-0.0158	0.0000	0.0000	-0.067	-0.557
Pool 8						
1	-0.0051	-0.0183	0.0101	0.0029	-0.056	-0.645
2	-0.0051	-0.0183	0.0065	0.0059	-0.048	-0.432
3	-0.0051	-0.0183	0.0046	-0.0013	-0.079	-0.687
4	-0.0051	-0.0183	0.0047	-0.0011	-0.078	-0.679
5	-0.0051	-0.0183	0.0028	-0.0041	-0.092	-0.786
6	-0.0051	-0.0183	0.0000	0.0000	-0.078	-0.645
Pool 9						
1	-0.0005	-0.0103	0.0018	0.0003	-0.038	-0.349
2	-0.0005	-0.0103	0.0007	0.0029	-0.029	-0.258
3	-0.0005	-0.0103	0.0009	0.0019	-0.033	-0.293
4	-0.0005	-0.0103	0.0002	-0.0019	-0.049	-0.427
5	-0.0005	-0.0103	-0.0009	-0.0024	-0.052	-0.445
6	-0.0005	-0.0103	0.0000	0.0000	-0.042	-0.361

The final two columns are predicted residual effect values for two selected positions from the section—one position near the beginning of the section, and the last position of the section. The predicted values were calculated using:

$$\hat{y}_{rm} = b_0 + b_1(P) + b_2(\theta_1) + b_3(\theta_2) + b_4(\theta_3) + b_5(\theta_4) + b_6(\theta_5) + b_7(P\theta_1) + b_8(P\theta_2) + b_9(P\theta_3) + b_{10}(P\theta_4) + b_{11}(P\theta_5). \quad (2)$$

where P represents the item's serial position in the test and $P\theta$ is the interaction of item-position-by-θ. For example, to calculate the predicted residual value for all test takers in the lowest ability category (the θ_1 category) for position 30 on the verbal section, all of the terms that contain $\theta_2 - \theta_5$ can be removed from the model, and the equation becomes

$$\hat{y}_{rm} = b_0 + b_1(P) + b_2(\theta_1) + b_7(P\theta_1). \quad (3)$$

For θ_6, all of the terms containing θ drop out of the model.

Both the sign and magnitude of the predicted residual values are noteworthy. Negative values indicate that test takers in a particular ability group took less time to answer items in that position; positive values indicate that they took more. Small values indicate that much of the variability of the response times to all items seen in that position was explained by the item and test-taker effects; the more extreme residual values indicate that there is consistent variation over and above the item and test-taker effects. If a predicted residual value is large and negative, it is an indication that, in that position, test takers in that ability group are responding more quickly to items than expected given the test-taker and item averages. The examination of predicted residual effects by θ category reveals how the prediction of the residual effect by position varies for different groups of test takers.

Tables 10 and 11 show the coefficients and predicted values for the verbal sections on Datasets 1 and 2. For all the pools in both datasets, it is apparent that for a position early in the section, the predicted residuals are close to zero. The predicted residuals are large and negative in the last position in the section for the low scorers, and approach zero as the θ category increases. This decrease in residual effect as θ increases is consistent with the results from the test-taker regressions: Test takers who do well on the verbal section work more quickly, and so their response times are not affected very much by speededness as they near the end of the section. The low-scoring test takers, on the other hand, work more slowly; by the end of the section, some respond very quickly to the items, indicating that the section may be speeded for low-ability test takers.

The coefficients and predicted residuals for the quantitative section are presented in Tables 12 and 13. The predicted values remain close to zero for a position near the beginning of the section. In the last position of the section, the predicted residuals do not consistently decrease as in the verbal section—in fact, for Pools 11, 12, and 13, the values are larger for the higher-θ test takers. This effect is even more pronounced in the Dataset 2 pools. On this section, it is the high scorers who give speeded responses by the end of the section. This is consistent with the item effect regression results. The more difficult quantitative items take longer to do, and so it is the test takers who are receiving these items who are running out of time near the end of the section.

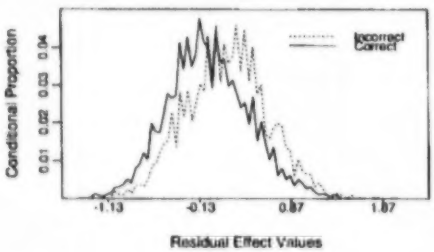
The analytical coefficients and predicted values are presented in Tables 14 and 15. The predicted values differ greatly between position 4 and position 35. Near the beginning of the section, no group of test takers has large predicted residuals. By the end of the section, all of the test-taker groups have large negative residuals. These results are consistent with the earlier finding that the analytical section is very speeded. Earlier results also showed that more difficult analytical items may take longer to complete, and that higher ability test takers may be working more quickly on analytical items; it is possible that those two effects cancel, and so the high-scoring test takers tend to have predicted residuals that are as large as the ones for the other test takers.

Distributional Plots

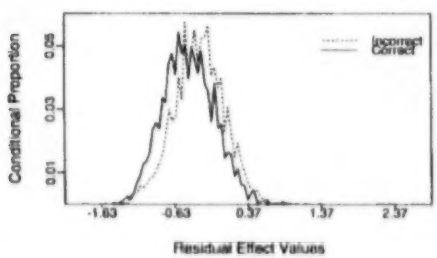
Selected graphs that show the distribution of the residual response times for discrete items in early, middle, and late serial positions are presented in Figures 1A–6B (verbal), Figures 7A–12B (quantitative), and Figures 13A–18B (analytical reasoning). Distributions for test takers who responded correctly are shown with a solid line, and those with incorrect responses are shown with a dashed line. Conditional proportion is

on the vertical axis, and the residual effect values—the log response times for test takers to the item each saw in that position, with the item and test-taker effects already removed—are on the horizontal axis. The proportions are conditional on whether the response was correct or incorrect; within each distribution, the proportions sum to one. Means, standard deviations, and sample sizes are provided for both correct (C_M, C_{SD}, and C_N) and incorrect (I_M, I_{SD}, and I_N) responses for each figure. If rapid-guessing behavior is present for a position, it appears as large, negative residuals that are more likely to be incorrect than correct.

The first six figures present the graphs for the verbal sections. The graphs for the items in the early serial positions are very similar across datasets (Figures 1A–2B). The distributions for both incorrect and correct responses are approximately normal. The means for the correct responses are close to zero, while the means for the incorrect responses are higher (from 0.15 to 0.20); the standard deviations for both distributions are the same. The graphs for positions near the middle of the verbal section (Figures 3A–4B) are very similar to the graphs for the earlier positions. The correct and incorrect distributions in these figures support the test-taker regression conclusions; test takers who know the answers respond more quickly in these positions than test takers who do not know the answers.

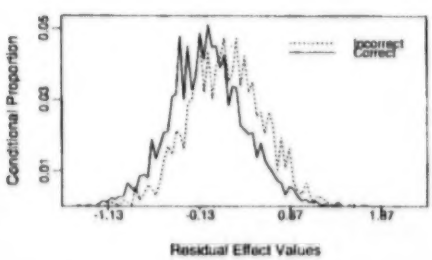


Position 5 C_M = -0.03 C_{SD} = 0.45 C_N = 3746 I_M = 0.17 I_{SD} = 0.45 I_N = 2762

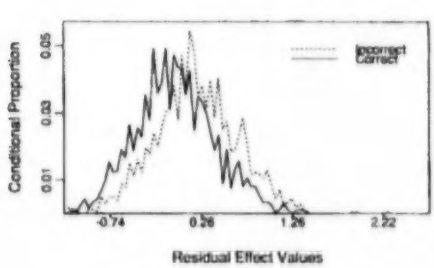


Position 6 C_M = -0.02 C_{SD} = 0.44 C_N = 3526 I_M = 0.15 I_{SD} = 0.44 I_N = 2607

FIGURES 1A and 1B. *The PDF graphs for residual response times, for positions 5 and 6, verbal, Dataset 1*

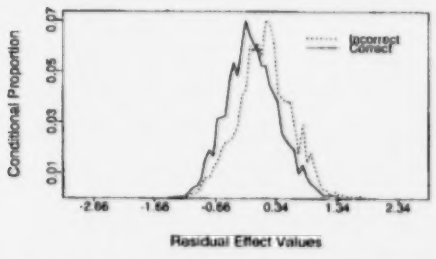


Position 5 C_M = -0.04 C_{SD} = 0.42 C_N = 2467 I_M = 0.16 I_{SD} = 0.42 I_N = 1652

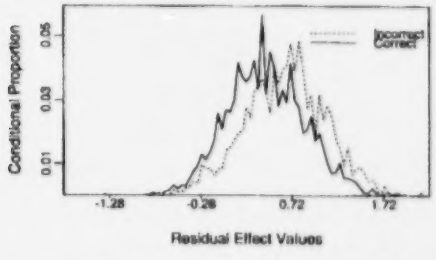


Position 6 C_M = -0.03 C_{SD} = 0.43 C_N = 2340 I_M = 0.20 I_{SD} = 0.45 I_N = 1780

FIGURES 2A and 2B. *The PDF graphs for residual response times, for positions 5 and 6, verbal, Dataset 2*

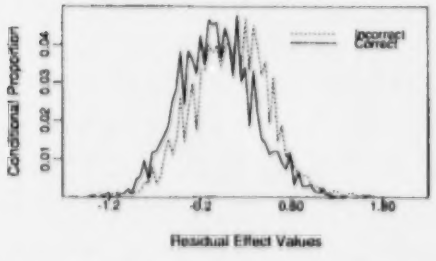


Position 14 C_M = -0.06 C_{SD} = 0.43 C_N = 4071 I_M = 0.14 I_{SD} = 0.45 I_N = 2104

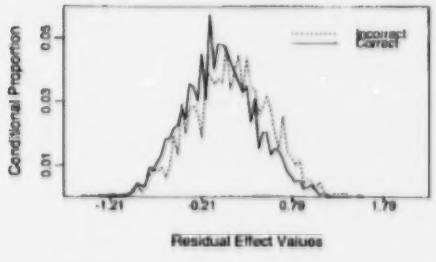


Position 15 C_M = -0.06 C_{SD} = 0.43 C_N = 3016 I_M = 0.14 I_{SD} = 0.45 I_N = 1763

FIGURES 3A and 3B. *The PDF graphs for residual response times, for positions 14 and 15, verbal, Dataset 1*



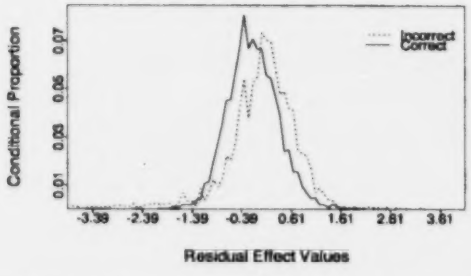
Position 15 C_M = -0.02 C_{SD} = 0.41 C_N = 2408 I_M = 0.13 I_{SD} = 0.42 I_N = 1707



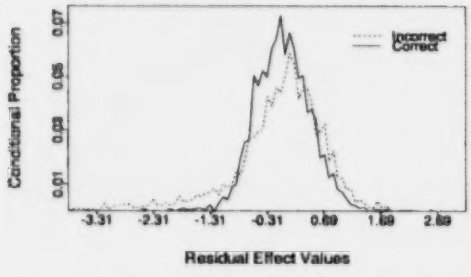
Position 16 C_M = -0.006 C_{SD} = 0.40 C_N = 2621 I_M = 0.11 I_{SD} = 0.43 I_N = 1499

FIGURES 4A and 4B. *The PDF graphs for residual response times, for positions 15 and 16, verbal, Dataset 2*

It is for the positions near the end of the section (Figures 5A–6B) that a difference is apparent. Although the majority of the incorrect responses continues to have more extreme positive residual values than the majority of the correct responses, a group of incorrect responses with large negative residuals now emerges. In each dataset, these rapidly guessed responses are most numerous for the final position in the section. The graphs show that by the final item position, there are responses that are very quick and not explained by the item content; these results combine with those of the residual regressions to suggest that the verbal section may be differentially speeded.

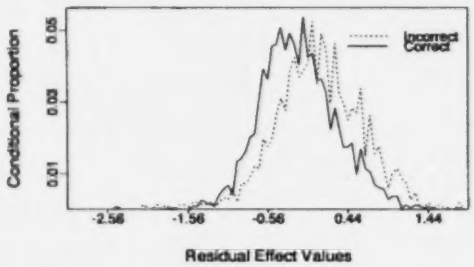


Position 28 C_M = -0.16 C_{SD} = 0.51 C_N = 3067 I_M = 0.009 I_{SD} = 0.77 I_N = 1711

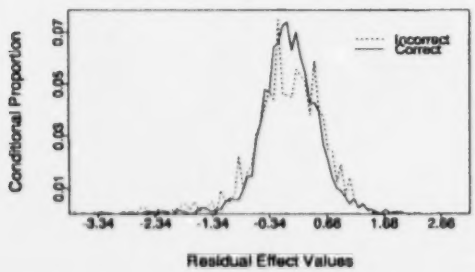


Position 30 C_M = -0.03 C_{SD} = 0.49 C_N = 2831 I_M = -0.11 I_{SD} = 0.84 I_N = 1773

FIGURES 5A and 5B. *The PDF graphs for residual response times, for positions 28 and 30, verbal, Dataset 1*



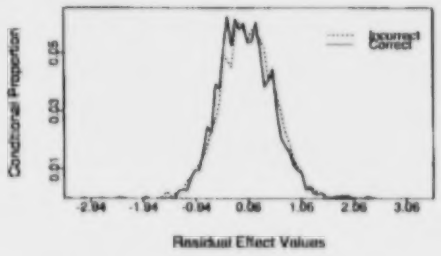
Position 28 $C_M = -0.17$ $C_{SD} = 0.48$ $C_N = 2477$ $I_M = 0.07$ $I_{SD} = 0.55$ $I_N = 1485$



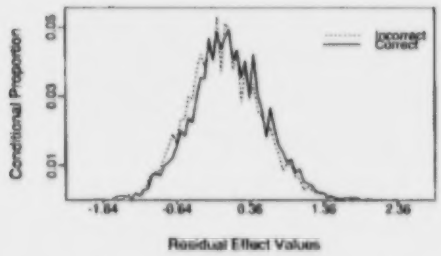
Position 30 $C_M = -0.01$ $C_{SD} = 0.55$ $C_N = 2339$ $I_M = 0.03$ $I_{SD} = 0.66$ $I_N = 1203$

FIGURES 6A and 6B. *The PDF graphs for residual response times, for positions 28 and 30, verbal, Dataset 2*

Figures 7A–12B concern the quantitative section residuals. The graphs of early and middle positions in the section all appear similar. The incorrect distributions do not appear to differ from the correct distributions in any way. The means of all of the distributions are very close to zero, and all of the standard deviations are equal or nearly so, and slightly larger than for the verbal section. Rapidly guessed responses are visible in the positions late in the quantitative section, and these effects are consistent across the two datasets. As with the verbal section, speeded responses that were not detected with the use of speededness statistics (because these positions were reached by most of the test takers) become apparent, and it is obvious that some test takers are randomly hitting answer keys by the time they reach late positions. Given the residual regression results for this section, it is possible that these rapid responders are the higher ability test takers.

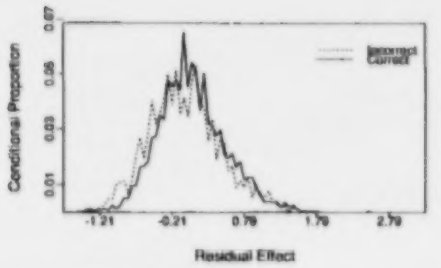


Position 3 $C_M = 0.007$ $C_{SD} = 0.52$ $C_N = 3636$ $I_M = 0.04$ $I_{SD} = 0.54$ $I_N = 3030$

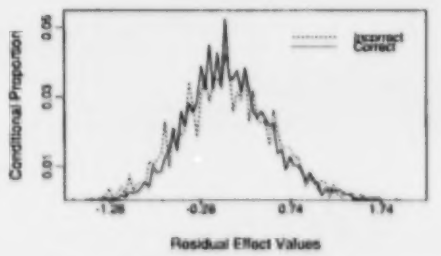


Position 6 $C_M = 0.07$ $C_{SD} = 0.51$ $C_N = 3295$ $I_M = 0.008$ $I_{SD} = 0.50$ $I_N = 3668$

FIGURES 7A and 7B. *The PDF graphs for residual response times, for positions 3 and 6, quantitative, Dataset 1*

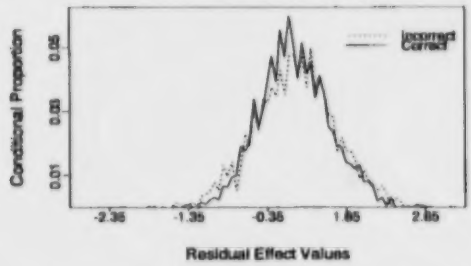


Position 4 $C_M = 0.06$ $C_{SD} = 0.47$ $C_N = 2448$ $I_M = -0.03$ $I_{SD} = 0.52$ $I_N = 1665$

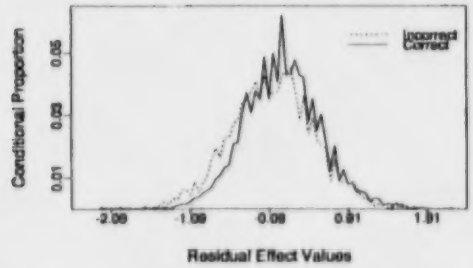


Position 6 $C_M = 0.05$ $C_{SD} = 0.47$ $C_N = 2364$ $I_M = 0.03$ $I_{SD} = 0.53$ $I_N = 1749$

FIGURES 8A and 8B. *The PDF graphs for residual response times, for positions 4 and 6, quantitative, Dataset 2*

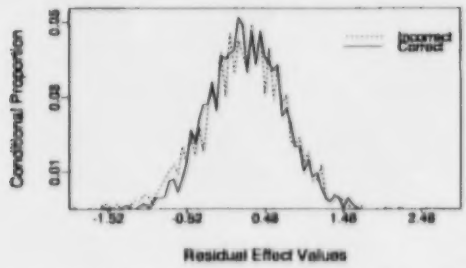


Position 14 C_M = -0.001 C_{SD} = 0.45 C_N = 2826 I_M = 0.002 I_{SD} = 0.51 I_N = 1867

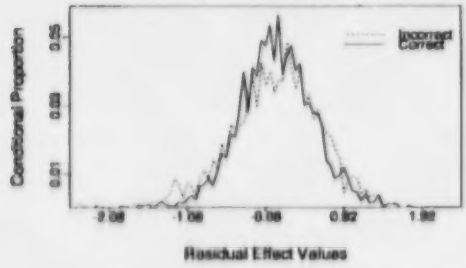


Position 15 C_M = 0.08 C_{SD} = 0.49 C_N = 3304 I_M = -0.05 I_{SD} = 0.55 I_N = 3177

FIGURES 9A and 9B. *The PDF graphs for residual response times, for positions 14 and 15, quantitative, Dataset 1*

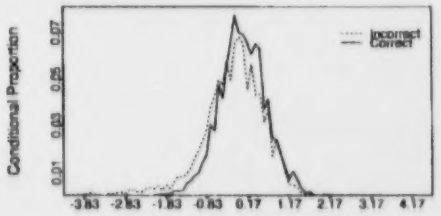


Position 14 C_M = 0.06 C_{SD} = 0.47 C_N = 1970 I_M = 0.03 I_{SD} = 0.53 I_N = 2018

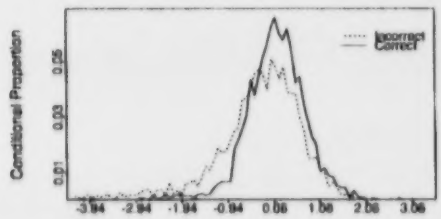


Position 15 C_M = 0.06 C_{SD} = 0.48 C_N = 2463 I_M = 0.03 I_{SD} = 0.58 I_N = 1645

FIGURES 10A and 10B. *The PDF graphs for residual response times, for positions 14 and 15, quantitative, Dataset 2*

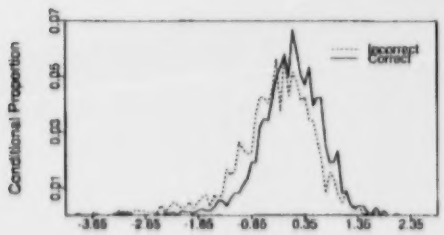


Position 26 $C_M = 0.03$ $C_{SD} = 0.60$ $C_N = 3411$ $I_M = -0.19$ $I_{SD} = 0.80$ $I_N = 3056$

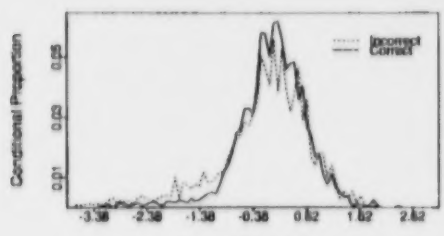


Position 28 $C_M = -0.10$ $C_{SD} = 0.65$ $C_N = 2718$ $I_M = -0.26$ $I_{SD} = 0.95$ $I_N = 2888$

FIGURES 11A and 11B. *The PDF graphs for residual response times, for positions 26 and 28, quantitative, Dataset 1*



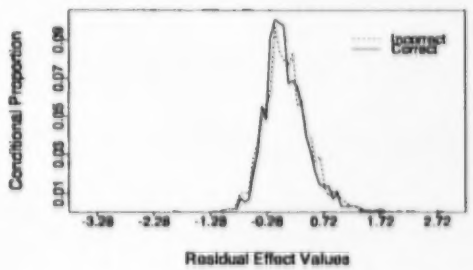
Position 26 $C_M = 0.09$ $C_{SD} = 0.61$ $C_N = 1725$ $I_M = -0.21$ $I_{SD} = 0.75$ $I_N = 1860$



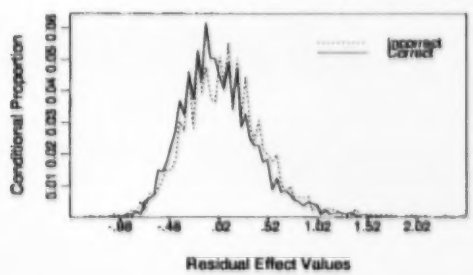
Position 28 $C_M = -0.03$ $C_{SD} = 0.67$ $C_N = 1949$ $I_M = -0.18$ $I_{SD} = 0.89$ $I_N = 1274$

FIGURES 12A and 12B. *The PDF graphs for residual response times, for positions 26 and 28, quantitative, Dataset 2*

The final six graphs (Figures 13A–18B) are for the analytical reasoning section. The figures for the early item positions show no differences between the distributions for the incorrect vs. correct responses. The means for the distributions are close to zero for Dataset 1 and increase slightly in Dataset 2. The standard deviations for all of the distributions are close to .40, which is smaller than for the verbal and quantitative sections. The PDF graphs for the item serial positions near the middle of the section are essentially the same as the graphs for the earlier positions, although there is a slight tendency for those who answer the item incorrectly to have a larger positive residual value. This indicates that those who have trouble with the items now take longer to complete them.

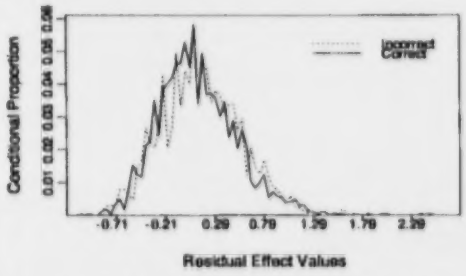


Position 6 $C_M = 0.06$ $C_{SD} = 0.40$ $C_N = 1738$ $I_M = 0.09$ $I_{SD} = 0.42$ $I_N = 1586$



Position 7 $C_M = 0.02$ $C_{SD} = 0.39$ $C_N = 3335$ $I_M = 0.09$ $I_{SD} = 0.42$ $I_N = 2742$

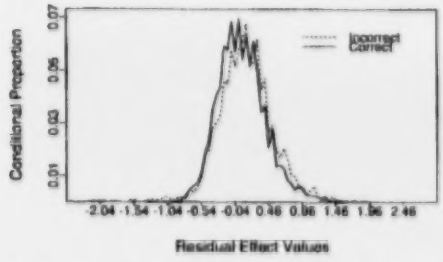
FIGURES 13A and 13B. *The PDF graphs for residual response times, for positions 6 and 7, analytical, Dataset 1*



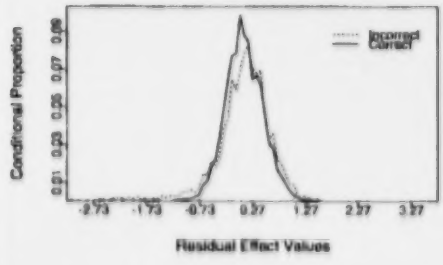
Position 7 $C_M = 0.12$ $C_{SD} = 0.40$ $C_N = 2424$ $I_M = 0.15$ $I_{SD} = 0.41$ $I_N = 1656$

Position 8 $C_M = 0.08$ $C_{SD} = 0.38$ $C_N = 2345$ $I_M = 0.13$ $I_{SD} = 0.43$ $I_N = 1734$

FIGURES 14A and 14B. *The PDF graphs for residual response times, for positions 7 and 8, analytical, Dataset 2*

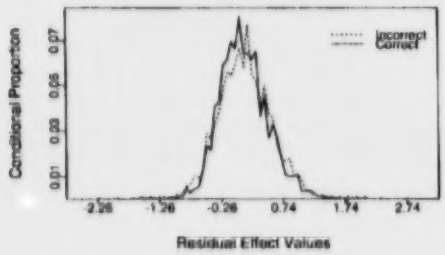


Position 18 $C_M = 0.06$ $C_{SD} = 0.37$ $C_N = 3233$ $I_M = 0.11$ $I_{SD} = 0.43$ $I_N = 2251$

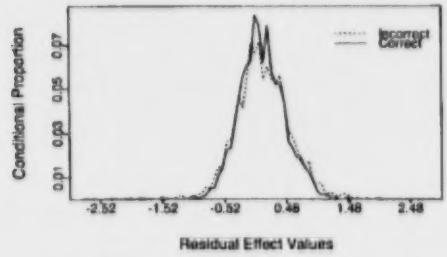


Position 23 $C_M = 0.11$ $C_{SD} = 0.38$ $C_N = 3446$ $I_M = 0.08$ $I_{SD} = 0.55$ $I_N = 1914$

FIGURES 15A and 15B. *The PDF graphs for residual response times, for positions 18 and 23, analytical, Dataset 1*

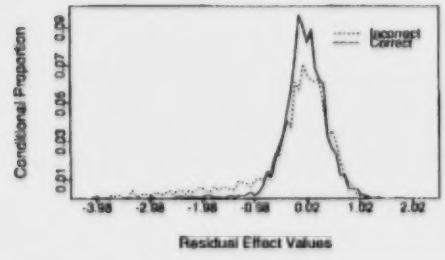


Position 14 $C_M = 0.10$ $C_{SD} = 0.39$ $C_N = 2457$ $I_M = 0.12$ $I_{SD} = 0.46$ $I_N = 1620$

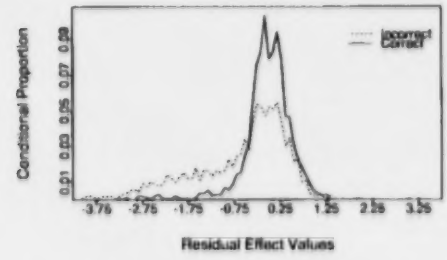


Position 19 $C_M = 0.08$ $C_{SD} = 0.39$ $C_N = 2295$ $I_M = 0.09$ $I_{SD} = 0.46$ $I_N = 1773$

FIGURES 16A and 16B. *The PDF graphs for residual response times, for positions 14 and 19, analytical, Dataset 2*

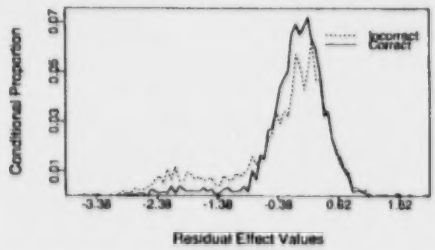


Position 30 $C_M = -0.04$ $C_{SD} = 0.47$ $C_N = 2344$ $I_M = -0.29$ $I_{SD} = 0.88$ $I_N = 2817$

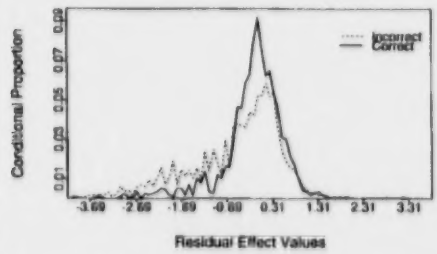


Position 35 $C_M = -0.02$ $C_{SD} = 0.55$ $C_N = 1772$ $I_M = -0.60$ $I_{SD} = 1.03$ $I_N = 1903$

FIGURES 17A and 17B. *The PDF graphs for residual response times, for positions 30 and 35, analytical, Dataset 1*



Position 30 $C_M = -0.09$ $C_{SD} = 0.57$ $C_N = 1827$ $I_M = -0.34$ $I_{SD} = 0.85$ $I_N = 1959$



Position 35 $C_M = -0.14$ $C_{SD} = 0.72$ $C_N = 1156$ $I_M = -0.49$ $I_{SD} = 0.99$ $I_N = 1158$

FIGURES 18A and 18B. *The PDF graphs for residual response times, for positions 30 and 35, analytical, Dataset 2*

The distributions for the late items positions show a large difference between the distributions for correct and incorrect responses. Residual times for those test takers who responded correctly to the items in these positions have means around zero and standard deviations only slightly higher than for the earlier positions. These late-position distributions look similar to the distributions for correct responses in earlier positions, although there is a small portion of the test takers with large negative residuals. In contrast, the distributions for the incorrect answers have much more extreme and negative means than do the correct distributions, and the proportions of test takers with these values are much higher than in any of the other positions. In fact, for the final position (Figure 17B for Dataset 1 and Figure 18B for Dataset 2), the distribution of incorrect responses appears to be a mixture of two normal distributions—one that contains test takers whose residuals were the same as the correct responders, and one that has a negative mean and a high variance, containing a large number of rapid guessers.

Discussion

The focus of the current study is the examination of item-level response times on an operational CAT. The analyses consisted of item, test-taker, and residual regressions and the graphing of residual response times. The test is divided into three sections—verbal, quantitative, and analytical reasoning—that were each administered under two different scoring rules—an 80% rule (Dataset 1) and a proportional-adjustment rule (Dataset 2). All of the analyses were performed separately for each section and dataset (as well as by pool), to see if different results would appear for different sections and scoring rules. These analyses were exploratory, and produced a multitude of findings that can be organized for summary and discussion in several different ways—by sets of analyses, by test section, or by scoring rule. The analyses produced very different results for the different sections of the test, and so an understanding of these results may best be accomplished by discussing each test section in turn, consolidating the results from all the analyses within each section. The interpretation of any scoring-rule differences found within a test section can be discussed within each section.

The Verbal Section

The item regressions for this section of Dataset 1 showed no relationship between the component of response time due to item effects and the item parameters, but the Dataset 2 item regressions showed a small positive relationship of response time with item difficulty, suggesting that the more difficult verbal items may take longer to complete. However, that relationship is very weak compared to the results for the test-taker regressions, where θ is predictive of the test-taker component of time in both datasets: The higher the ability, the less time a test taker uses on the items. The regression of the residual response times on position, θ and the position-by-θ interaction for both datasets produced predicted residuals that are more negative for low scorers than for high scorers, meaning that the low scorers are more likely to produce speeded responses near the end of the section. The graphs of the distributions of residual response time show rapidly guessed responses in item serial positions near the end of the section; given the regression results, these rapid guessers are probably the lower scoring test takers. Thus, the assumption from earlier research that θ is independent of rapid-guessing behavior (Oshima, 1994; Schnipke, 1995; Schnipke & Scrams, 1996) is not supported by these results from the verbal section, and the overall conclusion on this test section is that those with more verbal ability work more quickly than those with less.

All of the findings for the verbal section are stronger in the proportional-adjustment rule dataset than in the 80%-rule dataset, and so a question arises: Why do the scoring rules for the two datasets have a different effect on how quickly the test takers respond to the items? If traditional speededness indices, such as percent of test takers completing the test or some subset of it, were calculated, the 80%-rule dataset would appear to be the more speeded of the two datasets. However, those values would reflect the fact that test takers who were not rushed at all may have stopped early because they were allowed to. In fact, the proportional-adjustment dataset is likely to be the more genuinely speeded one; it has the same time limit as the 80%-rule dataset, and it also assigns a penalty to anyone who does not answer all items. Essentially, in Dataset 2, the test takers needed to complete all 30 verbal items in 30 minutes to maximize their score, while in Dataset 1, they had to complete only 24 items.

If Dataset 2 is assumed to be more speeded than Dataset 1, then one explanation for the different findings could be that the test takers who are able to correctly answer the more difficult verbal items are also able to do so at a higher rate than test takers who can correctly answer only the less difficult items. When the section is not very speeded, such as when the 80% rule is in effect, there is not a large difference among test takers in percentage of items completed, and less of a relationship between ability and speed emerges. If the section is more speeded, the test takers with lower ability are more affected by that change, because they need more time to answer the items targeted to their ability level.

This potential difference in the effect of time limits across the ability levels has implications for differential speededness research. Earlier studies have shown that black and Hispanic subgroups do not reach the end of verbal sections as often as whites on several different standardized tests (Dorans et al., 1988; Lawrence, 1993; Schmitt et al., 1991). It is possible that these subgroups would show increased rapid-guessing behavior for items that they do reach. Grouping of test takers both by ethnic subgroup and by ability level would be useful in detecting whether it is all members of a particular subgroup or only those members at the extremes of ability who exhibit different amounts of rapid-guessing behavior.

The Quantitative Section

When the regression analyses are performed separately for the item and test-taker effects in Dataset 1, the item parameters do not predict much of the variability in the item effects, nor does ability appear to predict very much of the test-taker effect variability. In addition, the predicted residual values for this dataset do not appear to vary consistently across the ability levels, so no clear relationships of item and test-taker characteristics with response time effects are observed.

A relationship does appear when the proportional-adjustment rule is in use: The more difficult quantitative items take longer to solve. The predicted residual values from the residual regression models support this conclusion. The values from Dataset 2 are more negative on the last item for the high scorers than for the low scorers, indicating that the test takers with high quantitative ability, who received the more difficult items, are more likely than the other test takers to run out of time and resort to rapid-guessing behavior.

The interpretation of the item regression results is simple, on one hand: One explanation is that the more difficult quantitative items are problems that involve more steps than the easier ones, so that the work requires more time even if the test taker knows how to do it. On the other hand, explaining why this relationship does not appear in Dataset 1 but does appear in Dataset 2 is not as simple. Even if the assumption is correct that Dataset 2 is more speeded than Dataset 1, that assumption does not appear to explain why test takers would use more time to complete the difficult items under the more speeded circumstance than under the less speeded one.

One possible answer may be that, if difficult quantitative items tend to take longer to complete even for those who can answer them correctly, the higher ability test takers may know this and factor that difference into their response strategies. Assume that when the 80% rule is in effect, the higher ability test takers actively use the knowledge that they have to answer only 23 of the 28 section items. This means that they can take more time to answer each item, and so can spend time working out even the items that seem easy at first, just to make sure that they are not missing a tricky answer or being fooled by an attractive distractor. It may make more sense to a test taker to take the time to answer 80% of the items correctly, rather than rush and answer a few extra items incorrectly. If that was their strategy, their response speed might not be highly affected by the item difficulty.

In contrast, when the section becomes more speeded under the proportional-adjustment rule, the higher ability test takers know that, even if they answer every item that they reach correctly, their score will be adjusted downwards when they do not answer every item. This may result in a change of strategy such that the items that appear easy are assumed to be easy, and little time is wasted on them. The test takers instead use their time on the items that are more difficult. This would result in a more pronounced relationship between response time and item difficulty in the more speeded situation. This is, however, only one potential explanation of the results.

The Analytical Reasoning Section

The results from the item and test-taker regressions suggest that, on the analytical section, more difficult items may take longer to complete, and the higher ability test takers may be working more quickly. The relationship between item difficulty and the item effect of response time is stronger than the relationship between ability and test-taker effect, and by the time the last item is reached, large negative residuals are predicted for all the test takers. The high-scoring test takers do not appear to have a time advantage and seem to be as speeded as the low scorers by the end of this section. The graphs of the distributions of the residual component of response time for this section show the highest percentage of test takers with rapid-guessing behavior by the end of this section, relative to the other sections; given residual regression results it is likely that the rapid-guessing group includes test takers at all ability levels.

The regression results for the analytical section are more pronounced when the proportional-adjustment rule is in effect. It is possible that the interpretations from both the verbal and the quantitative sections are applicable to the analytical section. When the section becomes more speeded, the higher ability test takers are more able to speed up without sacrificing accuracy than the lower ability test takers, and so the relationship between ability and time is stronger. Also, in the more speeded situation, the test takers may decide to take longer on the items that seem difficult, and perhaps take too little time on the items that appear easy. The results from this section make it clear that the design for the analytical section on this CAT is a long way from Eignor, et. al.'s (1993) ideal of the unspeeded CAT.

Additional Topics

All of the conclusions from the regression analyses must be qualified by one fact: A great deal of the variability in response time was not explained by the variables in this study. It is possible that some important components of a test taker's response time to an item are predictable by variables that were not included in the analyses or variables that were not measured with these datasets, such as indicators for various personality traits (need for closure, impulsitivity, etc.). Presumably the high-stakes environment of the test motivated all of the test takers, but it is possible that there were test takers not classified as outliers who still had very low total section times (and probably very low scores as well). Motivation may well affect response times, and its level may differ greatly even in high-stakes testing situations.

The differing presence of speededness provided the opportunity to examine whether graphs of response times, with the item and test-taker effects removed, could be applied to CAT data in order to detect rapid-guessing behavior. These graphs show a distribution of rapidly guessed responses on each section of the test, and they appear useful for detecting speeded responses on CATs, despite the fact that different test takers saw different items in each position. One future change that could be made is to divide test takers by ability level before plotting residual response times; given the current findings, it is possible that test takers with different levels of ability will appear in the rapid-guessing groups for different sections of a test.

One last comment is needed, and that concerns the generalizability of these results. Due to the design of this CAT, only discrete response times were meaningful enough to be used in the fitting of the prediction models. Many LSAT items, including all of the reading comprehension and analytical reasoning items, are not discrete. The only results of the current study which generalize to the LSAT are the discrete analytical reasoning items, which generalize to the discrete logical leasoning items. A CAT design is needed which allows researchers to separate the response time needed for the initial reading of the passage from the

response time needed for the accompanying items, although some passage review time may be included in all of the items. One solution may be to remove some estimate of average reading time from the passage-based item response times before using response time as a dependent variable.

References

Dorans, N. J., Schmitt, N. J., & Bleistein, C. A. (1988). *The standardization approach to assessing differential speededness* (Report No. ETS-RR 88-31). Princeton, NJ: Educational Testing Service.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation* (Report No. ETS-RR 93-56). Princeton, NJ: Educational Testing Service.

Godfrey, K. (1985). Fitting by organized comparisons: The square combining table method. In Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (eds.), *Exploring data tables, trends and shapes* (pp. 37–66). New York: John Wiley & Sons.

Lawrence, I. M. (1993). *The effect of test speededness on subgroup performance* (Report No. ETS-RR 93-49). Princeton, NJ: Educational Testing Service.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200–219.

Pashley, P. J. (1995, March). LSAT computerized assessment research agenda. *LSAC Research Proposal*. Newtown, PA: Law School Admission Council.

Schaeffer, G., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). *Field test of a computer-bases GRE general test* (Report No. ETS-RR-93-07). Princeton, NJ: Educational Testing Service.

Schaeffer, G., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE General Test* (Report No. ETS-RR-95-20). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T., (1991). *Differential speededness and item omit patterns on the SAT* (Report No. ETS-RR-91-50). Princeton, NJ: Educational Testing Service.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Schnipke, D. L., & Pashley, P. J. (1997, March). *Assessing subgroup differences in item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Schnipke, D. L., & Scrams, D. J. (1996, June). *Modeling response times in testing with a two-state mixture model: A new approach to detect speededness*. Paper presented at the annual meeting of the Psychometric Society, Banff, Alberta, Canada.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213–232.

Scrams, D. J., & Schnipke, D. L. (1997). *Empirical response-time distribution functions: How should response-time information be represented in item banks?* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Slater, S. C., & Schaeffer, G. (1996, April). *Computing scores for incomplete GRE general computer adaptive tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Swanson, D. B., Featherman, C., Case, S. M., Luecht, R. M., & Nungester, R. J. (1997, March). *Relationship of response latency to test design, test-taker proficiency, and item difficulty in computer-based test administration.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Thissen, D. M. (1983). Timed testing: An approach using item response theory. In D. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). New York: Academic Press.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (2002). *Using response-time constraints in item selection to control for differential speededness in computerized adaptive testing* (Computerized Testing Report 98-03). Newtown, PA: Law School Admission Council.

Appendix

Description and Example of the SCT Method on a Two-Way Table

Table A1 provides a 4-test taker by 3-item matrix, with response time in seconds as the data values. To calculate effects for the items (columns), set up a matrix that lists each item pair. List the difference between the two elements of the item pair for each test taker (or row) as shown in Table A2.

TABLE A1
Response times of four test takers to three items

Test Taker	Item 1	Item 2	Item 3
1	1	2	3
2	2	3	4
3	4	5	6
4	1	2	3

TABLE A2
Differences between each pair of item response times, presented for each test taker

	Item (Column) Pair		
Test Taker	1,2	1,3	2,3
1	-1	-2	-1
2	-1	-2	-1
3	-1	-2	-1
4	-1	-2	-1
Mean	-1	-2	-1

These data were created so that the matrix in Table A2 is perfectly additive, so each row has the same value for each column pair; another way to say this is that each person took one second longer on item 2 than item 1. Next, take the mean for each column pair. In the article by Godfrey (1985), the median of the column pairs across the rows is computed, but the mean, which is computationally easier to do with large matrices, may be used as well.

Next, take these means and place them into an anti-symmetric square combining table, shown in Table A3, with the opposite sign of each value in the upper triangle, and zeroes on the diagonals (If the mean difference between item 1 and item 2 is -1, then the mean difference between item 2 and item 1 is 1). Finally, take the means of the columns of this anti-symmetric matrix; these are the item effects with test-taker effects already removed.

TABLE A3
The square combining table for the item effects

	Items		
Items	1	2	3
1	0	1	2
2	-1	0	1
3	-2	-1	0
Mean	-1	0	1

The means listed in Table A3 are the item effect vector v. The same vector could have been produced for this data by taking the test-taker means of Table A1, subtracting them from each row, and then taking the item (column) means. To get test-taker effects (the vector w), the same process described above can be repeated for the rows of Table A1: calculate the differences between each test-taker pair, taking the means across item, place these means in a 4 by 4 test-taker antisymmetric matrix, and obtain the mean vector for that table. These resulting tables are shown in Table A4 and Table A5.

TABLE A4
Differences between each pair of test-taker response times, presented for each item

	Test-taker (Row) Pair					
Item	1,2	1,3	1,4	2,3	2,4	3,4
1	-1	-3	0	-2	1	3
2	-1	-3	0	-2	1	3
3	-1	-3	0	-2	1	3
Mean	-1	-3	0	-2	1	3

TABLE A5
The square combining table for the test-taker effects

	Test Takers			
Test Takers	1	2	3	4
1	0	1	3	0
2	-1	0	2	1
3	-3	-2	0	-3
4	0	1	3	0
Mean	-2	0	2	-0.5

Note: The means from Table A5 comprise the vector *w*.

The SCT method is not perfectly resistant to sparseness. In particular, if there are holes in the anti-symmetric matrix (Tables A3 and A5, above), then the final mean computation will not necessarily accurately represent the effects. If Table A1 had values removed in it such that test takers 1 and 2 saw no items in common, the data could appear as in Table A6.

TABLE A6
Response times of four test takers to three items, with missing data

Test Taker	Item 1	Item 2	Item 3
1	---	2	3
2	2	---	---
3	4	5	6
4	1	2	3

For this modified data, the item differences would produce the same means, as shown in Table A7.

TABLE A7
Differences between each pair of item response times, with missing data

	Item (Column) Pair		
Test Taker	1,2	1,3	2,3
1	---	---	-1
2	---	---	---
3	-1	-2	-1
4	-1	-2	-1
Mean	-1	-2	-1

The resulting square combining table for the item effects would still reproduce the full-data-matrix effects, as shown in Table A8

TABLE A8
The square combining table for the item effects, with data values removed

	Item	Item	Item
Items	1	2	3
1	0	1	2
2	-1	0	1
3	-2	-1	0
Mean	-1	0	1

The test-taker differences for Table A6, however, would fail to reproduce the means, as Table A9 makes clear.

TABLE A9
Differences between each pair of test-taker response times, with missing data

	Test-taker (Row) Pair					
Item	1,2	1,3	1,4	2,3	2,4	3,4
1	---	---	---	-2	1	3
2	---	-3	0	---	---	3
3	---	-3	0	---	---	3
Mean	---	-3	0	-2	1	3

The resulting square combining table for the item effects, shown in Table A10, would not reproduce the full-data-matrix estimate of w as shown in Table A5.

TABLE A10
The square combining table for the test-taker effects, with missing data

	Test Takers			
Test Takers	1	2	3	4
1	0	---	3	0
2	---	0	2	-1
3	-3	-2	0	-3
4	0	1	3	0
Mean	-1	-0.33	2	-0.5

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").